LLM4GKID: A Multimodal Large Language Model-driven Framework for Ghost Kitchen Identification

Weipeng Deng*, Yihong Tang*, Chaofan Wang, and Tianren Yang†

***Abstract*—The proliferation of ghost kitchens (i.e., delivery-only restaurants without physical storefronts) poses significant challenges for urban food system monitoring and regulatory oversight. These establishments maintain digital visibility on delivery platforms while eluding visibility in the physical public realm, creating information asymmetries that compromise transparency and consumer protection. This work presents LLM4GKID, a comprehensive approach for detecting ghost kitchens by leveraging the sensing capability of large language models to match Point of Interest records across platforms. The methodology integrates multiple information sources through a staged pipeline: geographic filtering to identify spatial candidates, language model-based semantic similarity assessment, visual consistency analysis of establishment imagery, and machine learning classification. Central to our contribution is the alignment failure detection mechanism, which systematically identifies delivery-only establishments lacking corresponding entries in crowdsourced review databases. Evaluation on a manually annotated dataset of restaurant POI pairs from Shenzhen, China, demonstrates substantial performance improvements over existing methods adapted for our task. The progressive filtering strategy significantly reduces computational complexity while maintaining high recall by conservatively selecting candidates. Our framework addresses fundamental challenges in category-specific business model detection, where traditional POI conflation approaches fail due to sparse category features and the spatial autocorrelation implied by Tobler's law. LLM4GKID overcomes these limitations through the integration of multimodal evidence and negative matching logic, enabling the accurate identification of establishments with asymmetric digital presence patterns. The framework allows downstream research in food access equity, regulatory compliance monitoring, and broader applications to other business models with asymmetric digital footprints. The source codes of LLM4GKID are available at https://github.com/weipengdeng/LLM4GKID.**

***Index Terms*—Ghost Kitchen, Large Language Model, POI conflation, Points of Interest, Multi-source Data Fusion, Platform Urbanism**

*Equal Contribution. †Corresponding Author.

W. Deng and C. Wang are with the Department of Urban Planning and Design, The University of Hong Kong, Hong Kong SAR, China, and also with the Urban System Institute, The University of Hong Kong, Hong Kong SAR, China, and also with the Shenzhen Institute of Research and Innovation, The University of Hong Kong, Shenzhen, Guangdong, China (e-mail: wpdeng@connect.hku.hk; chaofanw@connect.hku.hk).

Y. Tang is with the Department of Civil Engineering, McGill University, Montreal, Quebec, Canada (e-mail: yihong.tang@mail.mcgill.ca).

T. Yang is with the Department of Urban Planning and Design, The University of Hong Kong, Hong Kong SAR, China, and also with the Urban System Institute, The University of Hong Kong, Hong Kong SAR, China, and also with the Shenzhen Institute of Research and Innovation, The University of Hong Kong, Shenzhen, Guangdong, China (e-mail: tianren@hku.hk).

I. Introduction

The rapid expansion of online food delivery platforms has fundamentally reshaped urban foodscapes, facilitating the emergence of ghost kitchens (also known as dark kitchens and cloud kitchens) [1]. These delivery-only commercial cooking facilities operate exclusively through digital platforms without traditional storefronts [2], representing a paradigmatic shift from conventional restaurant models. Unlike traditional restaurants that rely on physical presence for brand visibility, ghost kitchens maintain complete operational invisibility while serving consumers through delivery apps [3], creating a parallel food economy that operates largely outside traditional regulatory frameworks.

This invisibility presents unprecedented challenges for food safety oversight, consumer protection, and urban governance. Traditional regulatory frameworks assume commercial food establishments operate within visible, accessible premises where health inspectors can conduct routine evaluations [4]. However, ghost kitchens fundamentally disrupt these assumptions by operating in concealed locations and maintaining no public-facing facilities for direct consumer interaction [5]. The proliferation of ghost kitchens has outpaced regulatory adaptation across major metropolitan areas worldwide, with studies indicating that 27% of restaurants on food delivery platforms in Brazil were classified as ghost kitchens [6]. Yet, they remain largely invisible in regulatory monitoring systems [3].

Deliberate obfuscation strategies and fundamental geographic principles together intensify the difficulty of detecting ghost kitchens. These establishments often operate in discreet or low-visibility settings to reduce costs and avoid regulatory scrutiny [2, 5], yet they still cluster with traditional restaurants because they benefit from similar locational advantages [7]. This spatial co-location is further shaped by Tobler's first law of geography, which states that "near things are more related than distant things." As a result, ghost kitchens and conventional restaurants often display highly similar spatial patterns in coordinate space despite having distinct operating models, making automated identification substantially more challenging [8].

Furthermore, the heterogeneity of data sources across food delivery platforms, mapping services, and business directories creates semantic inconsistencies that traditional POI (Point of Interest) conflation methods struggle to address [9, 10, 11, 12]. Ghost kitchens may appear as multiple distinct entities

on delivery platforms, while operating from a single location, or may be absent entirely from traditional business directories, creating systematic information asymmetries that compromise transparency and fair competition.

Large Language Models (LLMs) offer unprecedented capabilities for addressing these complex detection challenges. Unlike traditional semantic matching approaches that rely on token-based similarity or fixed embeddings, LLMs possess vast world knowledge and sophisticated contextual reasoning abilities that can interpret subtle naming variations, brand aliases, and creative obfuscation strategies employed by ghost kitchen operators. Recent advances in LLM applications to spatial data science demonstrate their potential for geographic reasoning and entity resolution tasks [10], yet their application to category-specific business model detection remains largely unexplored. The key insight is that LLMs can effectively leverage multimodal evidence integration and negative matching logic to identify establishments with asymmetric digital presence patterns—precisely the characteristic that defines ghost kitchens.

To address these multifaceted challenges, we introduce LLM4GKID, a comprehensive framework that leverages LLM-powered alignment failure detection to systematically identify ghost kitchens by matching POI records between online delivery platforms and crowdsourced review platforms. Unlike traditional POI conflation methods that focus on successful matches, our approach employs a cost-effective sequential pipeline that combines geographic filtering, LLM-assisted semantic similarity assessment to capture nuanced restaurant naming patterns and variations, computer vision analysis of establishment imagery, and supervised machine learning classification. Additionally, through cross-platform validation, LLM4GKID overcomes deliberate misinformation on a single platform (e.g., false "dine-in available" labels), ensuring robust detection. The key innovation lies in recognizing that ghost kitchens can be identified through their systematic absence from crowdsourced review platforms, thereby transforming the detection problem into a negative matching task where LLMs excel at disambiguating complex semantic relationships. We make three key contributions to computational social systems and urban analytics:

- **Technical Innovation:** We introduce LLM4GKID, a novel cross-platform POI conflation framework that integrates LLM-driven semantic understanding with spatial proximity filtering, logo-based visual consistency assessment, and XGBoost-driven classification to achieve superior accuracy (F1: 92%) in ghost kitchen identification compared to existing approaches.
- **Dataset Contribution:** We develop and publicly release a comprehensive benchmark dataset of manually validated restaurant POI pairs, establishing the first standardized testbed for cross-platform restaurant conflation and ghost kitchen detection research.
- **Societal Impact:** We demonstrate LLM4GKID's practical utility in uncovering hidden ghost kitchens within real-world commercial ecosystems, providing policymakers and urban planners with an LLM-powered scalable tool for monitoring the digital foodscape, ensuring consumer transparency, and informing strategic decisions around urban culinary infrastructure development.

II. Related Work

While digital platform economies have attracted increasing scholarly attention, the specific challenge of ghost kitchen identification remains relatively underexplored. Ghost kitchens, which operate exclusively through delivery while often presenting themselves as dine-in restaurants, embody a new urban phenomenon on driven by the integration of digital platforms, logistics networks, and physical retail systems [4]. Existing studies on platform economies have primarily focus on topics such as fraud detection, product classification, and delivery capacity predictions rather than store channel detection [13], yet few have directly addressed the detection and mapping of ghost kitchens as distinct entity.

Early research has relied on manual inspection, keyword-based filtering, and street-view imagery to detect ghost kitchens. These approaches provide useful clues but are difficult to scale, especially in areas lacking timely or comprehensive street-view coverage. Given the inherent data inconsistencies of ghost kitchen operations, multi-source data fusion has emerged as a more effective strategy, with POI conflation serving as the foundation for aligning cross-platform records [14]. Although recent advances in LLMs have improved general place alignment [15], existing methods still lack the negative matching logic required to detect these establishments. To address these challenges, we review prior studies in three domains: (1) existing approaches for ghost kitchen identification, (2) POI conflation and multimodal integration, and (3) semantic matching using traditional encoders and large language models.

A. Existing Approaches for Ghost Kitchen Identification

Research directly addressing ghost kitchen identification remains limited but has made valuable progress across several disciplines. Drawing on established definitions [1, 16], early studies identified ghost kitchens through manual inspection of platform listings and street-view verification [6, 17], providing the first systematic foundations for mapping delivery-only restaurants. A second stream of research uses keyword- and metadata-based filtering, identifying restaurants through descriptors like "Deliveroo Editions" or service-mode inconsistencies (e.g., dine-in vs. delivery-only) [18]. These approaches facilitate more efficient screening, but they depend on self-reported attributes and are easily misled by intentionally deceptive information [4], posing significant challenges for accurate, large-scale identification. Despite these contributions, existing methods remain largely heuristic and platform-dependent, limiting their ability to detect ghost kitchen effectively. These limitations highlight the need for a multimodal, multi-source framework that integrates spatial, semantic, and even visual signals to more robustly identify and verify ghost kitchens across digital platforms.

B. POI Conflation and Multimodal Integration

POI conflation offers a more systematic pathway for ghost kitchen identification by aligning inconsistent cross-platform records. These approaches have evolved from rule-based methods to sophisticated machine learning techniques, yet remain limited in multimodal integration. Early approaches relied on deterministic rules that combined name similarity and spatial distance thresholds [19], using either fixed or adaptive weighting schemes [20, 21]. While straightforward, these methods struggle with parameter generalization across datasets [14]. Supervised learning approaches compute multiple similarity features for candidate pairs, using classifiers to learn optimal combinations [22]. Traditional methods employed Logistic Regression and Random Forests [23, 24], while recent approaches leverage deep learning techniques including BERT-based semantic embeddings [25], knowledge graph [26], multi-view encoders [27], and sequence inference models [28]. These achieve high accuracies ($> 90\%$) but require substantial labeled training data [29, 30]. However, most research focuses primarily on textual and spatial features. Geographic distance serves as a universal filtering for scalability [27], while semantic approaches use string metrics, preprocessing, and embedding techniques [31, 32]. Some studies incorporate user-generated content [24, 8], but visual cues remain virtually absent despite digital platforms containing informative imagery [14]. This multimodal gap motivates our approach, which integrates multi-source spatial, semantic, and visual signals to detect ghost kitchens.

C. Semantic Matching: Large Language Models vs. Traditional Encoders

Another emerging opportunity is the application of large language models (LLMs) to content detection and urban sensing [33, 34]. Traditionally, the "semantic" aspect of POI conflation has been handled by relatively localized text processing, specifically comparing names or descriptions using token-based similarity or moderate-sized pre-trained embeddings. For example, studies up to 2022 often used encoders like Word2Vec or BERT to vectorize POI names and addresses [25], which already proved more robust than raw string matching. Yet, these models are limited to the information within the input text (e.g., the two names being compared) and their fixed, pre-trained knowledge. In contrast, generative LLMs contain a vast breadth of world knowledge and exhibit strong contextual reasoning abilities. Recent work in spatial data science has begun to explore LLMs. For instance, GeoGPT and GeoLLM use prompt-based queries on LLMs to answer geographic questions, and POI-Enhancer leverages an LLM to enrich POI feature representations with richer textual knowledge [35]. However, using LLMs for fine-grained POI conflation is still in its infancy. One challenge is how to extract useful, contextual semantic features from an LLM without heavy cost (e.g., an LLM could, in principle, infer that "Sunshine Burgers" and "Sunshine Diner" at the same address likely refer to the same business, but operationalizing this insight is nontrivial). Early studies on general entity resolution suggest that while prompting large models can yield impressive accuracy [36], smaller domain-tuned models often perform comparably. For example, [37] conducted a comprehensive evaluation of entity matching techniques and found that fine-tuned small language models (e.g., a BERT-based classifier) can achieve on-par performance with prompted LLMs, at a fraction of the deployment cost. This finding tempers the assumption that "bigger is always better". In many cases, carefully engineered local models or task-specific encoders can rival an off-the-shelf LLM on matching tasks. Nonetheless, LLMs bring certain advantages that remain underexploited in POI alignment. They can interpret context in free-form text (e.g., understanding that two restaurants' descriptions both mention the same signature dishes in their names). They can even resolve obscure name variations using real-world knowledge [36, 37]. However, limited work has explored LLM reasoning in POI conflation for scenarios involving entities that deliberately obscure their operational characteristics. LLM4GKID addresses this gap by leveraging LLM-level semantic and visual sensing capability to detect subtle inconsistencies and creative aliasing strategies that traditional approaches would overlook, enabling more robust identification of deliberately deceptive operational models.

III. Methodology

This section introduces LLM4GKID, a multimodal framework for aligning restaurant POIs across heterogeneous online and offline platforms, to identify *ghost kitchens*—delivery-only establishments lacking physical storefronts. LLM4GKID integrates *spatial*, *semantic*, and *visual* information from multiple data sources in a stepwise manner (Figure 1).

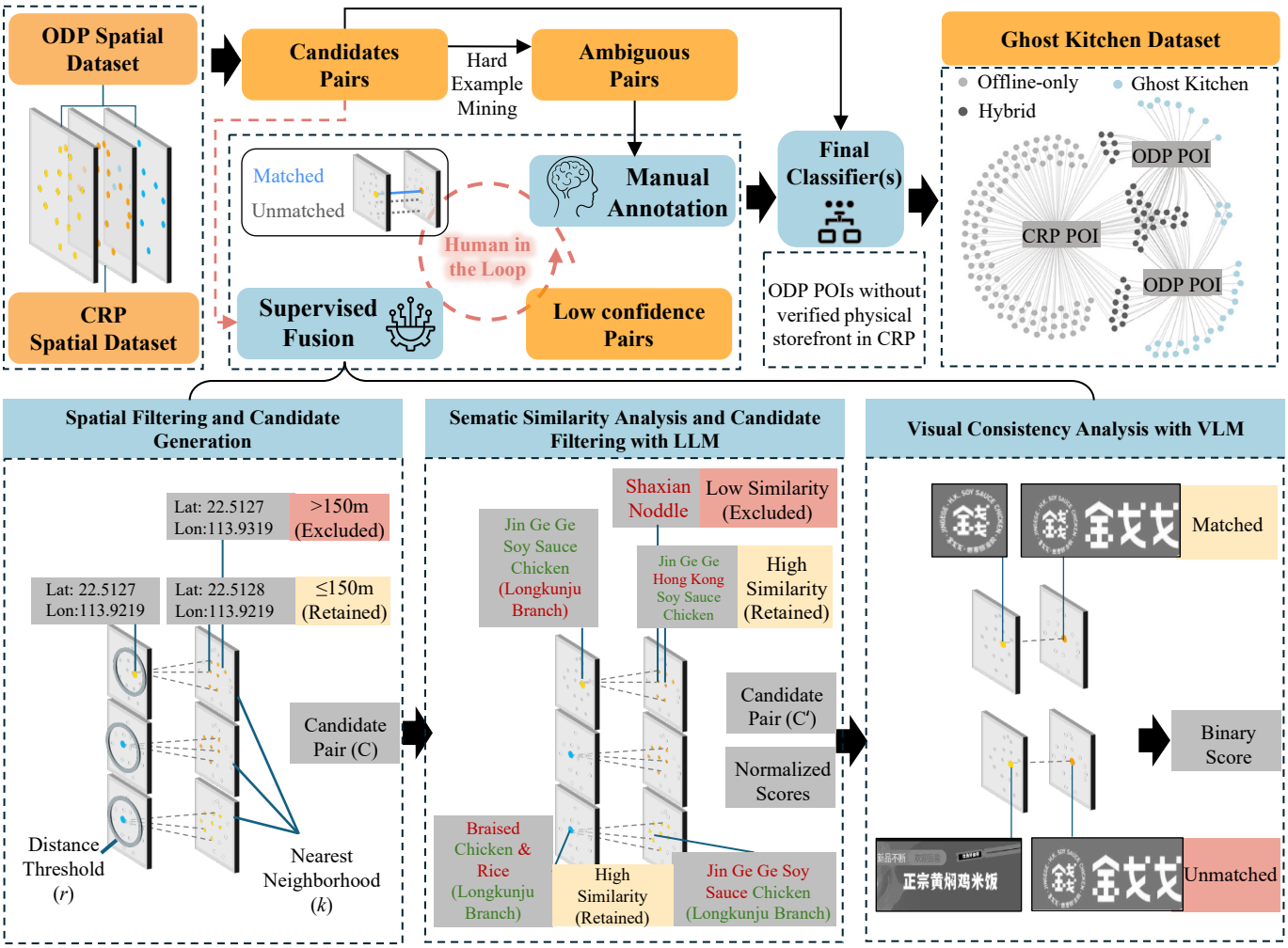
A. Overview

The core objective of LLM4GKID is to identify ghost kitchens by performing multimodal alignment of POIs across heterogeneous platforms. Specifically, we distinguish between two major types of sources: Online Delivery Platforms (ODPs), which provide digital listings of restaurants available for online ordering (e.g., food delivery platforms), and Crowdsourced Review Platform (CRP), which document venues with verified physical locations through user-generated reviews, check-ins, or on-site photos.

Formally, we consider two sets of POIs: one from an ODP $\mathcal{O} = \{o_i\}_{i=1}^{N}$, and one from a CRP $\mathcal{P} = \{p_j\}_{j=1}^{M}$. Each POI $x \in \mathcal{O} \cup \mathcal{P}$ is described by a multimodal tuple: $x = (x^{\text{geo}}, x^{\text{smt}}, x^{\text{vis}})$, where $x^{\text{geo}} \in \mathbb{R}^2$ denotes geographic coordinates (latitude and longitude), x^{smt} represents semantic attributes in textual form (e.g., name, address), and x^{vis} denotes associated visual information (e.g., storefront or logo images), if available. Our task is to determine whether a given ODP POI $o_i \in \mathcal{O}$ and a CRP POI $p_j \in \mathcal{P}$ refer to the same real-world restaurant entity. To this end, we define a binary matching function:

$$f(o_i, p_j) \to \{0, 1\}, \tag{1}$$

where $f(o_i, p_j) = 1$ indicates a match, and $f(o_i, p_j) = 0$ indicates that the two POIs correspond to different entities.



Notes: **1) ODP**: Online delivery platform that provides digital listings of restaurants available for online ordering; **2) CRP**: Crowdsourced review platform that documents venues with verified physical locations through user-generated reviews, check-ins, or on-site photos.

Fig. 1: The Architecture of the LLM4GKID Approach

To learn this function, we optimize a parameterized model f_θ that minimizes the empirical risk over a labeled training set $\mathcal{D} = \{(o_i, p_j, y_{ij})\}$, where $y_{ij} \in \{0, 1\}$ denotes the ground-truth alignment label:

$$\min_{\theta} \mathcal{L}(\theta) = \sum_{(o_i, p_j, y_{ij}) \in \mathcal{D}} \ell\big(f_\theta(o_i, p_j), y_{ij}\big), \tag{2}$$

where $\ell(\cdot, \cdot)$ is a suitable binary classification loss function, such as binary cross-entropy. The model f_θ jointly encodes the spatial, semantic, and visual features of POI pairs to infer their alignment likelihood.

The key insight of LLM4GKID is that ghost kitchens can be inferred from *alignment failure*. Drawing from the definition of ghost kitchens, restaurants that operate solely via online delivery without public-facing physical venues for dine-in services [1, 16], we identify such entities based on their absence from CRP datasets. Specifically, an ODP POI $o_i \in \mathcal{O}$ is classified as a ghost kitchen candidate if it fails to align with *any* CRP POI:

$$f(o_i, p_j) = 0 \quad \text{for all } p_j \in \mathcal{P}. \tag{3}$$

This formulation embeds ghost kitchen detection directly within the alignment framework. Rather than requiring a separate detection module, the absence of a match serves as implicit evidence of digital-only existence. The strength of this approach lies in its negative reasoning: ghost kitchens are identified not by what they explicitly possess, but by what they systematically lack, *i.e.*, user-verified physical presence across crowdsourced review platforms.

As shown in Figure 1, the LLM4GKID framework consists of three sequential components, each leveraging a different modality to improve alignment accuracy. The pipeline begins with spatial filtering, which efficiently narrows down the offline candidate set for each online POI using geographic distance thresholds. Next, we compute semantic similarity using language models to handle name variations and address inconsistencies. Visual consistency is then assessed via vision-language models (VLMs) based on storefront or logo imagery, providing an additional signal where available. Finally, these multimodal features are fused through supervised learning that outputs the final alignment decision. This stepwise design ensures cost-effectiveness by prioritizing lightweight fil-

ters before invoking more computationally expensive models, while also enabling robust matching through complementary evidence sources. In the following subsections, we provide a detailed description of each component.

B. Spatial Filtering and Candidate Generation

To efficiently reduce the alignment search space, LLM4GKID begins with spatial filtering based on geographic proximity. For each online POI $o_i \in \mathcal{O}$, we identify a candidate set of offline POIs $\mathcal{C}_i \subseteq \mathcal{P}$ consisting of those within a fixed radius τ_s, i.e.,

$$\mathcal{C}_i = \{p_j \in \mathcal{P} \mid h(o_i^{\text{geo}}, p_j^{\text{geo}}) \leq \tau_s\}, \tag{4}$$

where $h(\cdot,\cdot)$ denotes the *haversine distance* between two geographic coordinates. This spatial threshold τ_s is chosen based on domain knowledge and prior work in POI alignment [38], typically set between 100 and 200 meters to balance precision and recall. To efficiently compute all pairwise distances, we employ a BallTree-based nearest-neighbor search. It yields a reduced candidate set: $\mathcal{C} = \bigcup_{i=1}^{N} \{(o_i, p_j) \mid p_j \in \mathcal{C}_i\}$ of at most k spatially nearest neighbors per o_i, i.e., $|\mathcal{C}| \ll |\mathcal{O}| \times |\mathcal{P}|$.

This traditional spatial filtering step, also called *blocking* [27], serves as a lightweight and cost-effective preprocessing module. It eliminates implausible matches and drastically reduces the number of pairwise comparisons required in subsequent stages, which involve higher-complexity semantic and visual feature extraction. Importantly, it ensures that only geographically plausible POI pairs are passed to the multimodal alignment pipeline, thereby maintaining alignment accuracy while minimizing computational demand.

C. Semantic Similarity via Language Models

After spatial filtering, we assess the semantic similarity between each ODP-CRP POI pair in the candidate set $\mathcal{C}$ to refine alignment plausibility. Each POI $x \in \mathcal{O} \cup \mathcal{P}$ is associated with semantic attributes in textual form, which are used to evaluate the degree of semantic correspondence.

We first initialize a large language model (LLM) f_{LLM} with a fixed system prompt that encodes our matching criteria in a task-specific manner. Then, for each candidate pair $(o_i, p_j) \in \mathcal{C}$ we send only a lightweight user prompt, thereby minimizing token usage, and involve the model to compute a normalized similarity score via our prompt-base inference mechanism: $s^{\text{LLM}}(o_i, p_j) \in [0, 100]$.

In the actual implementation, only the candidates with the highest semantic similarity scores above τ_m would be retained in a refined candidate set $\mathcal{C}' \subseteq \mathcal{C}$. To maintain high recall and avoid false negatives that could hinder ghost kitchen detection, we adopt a conservative filtering strategy that prioritizes comprehensiveness over precision at this stage. While this approach still achieves an approximately over 90% reduction in the refined candidate set compared to the initial spatial filtering results, it deliberately retains pairs with ambiguous or borderline scores to ensure potentially valid matches are not prematurely discarded. This conservative threshold selection recognizes that false negatives at the filtering stage cannot be recovered by downstream classification, making recall preservation critical for reliable ghost kitchen identification. As a result, $\mathcal{C}'$ may include both accurate matches and semantically similar but incorrect pairs ("pseudo-matches"), which are intentionally preserved for final adjudication by the supervised classifier rather than being eliminated through aggressive early filtering.

For all pairs in $\mathcal{C}'$, we compute dense similarity features using a language model encoder. Let $f_{\text{LMEnc}}(\cdot)$ denote a shared text encoder applied to the semantic fields of o_i and p_j, resulting in vector representations $\mathbf{v}_i$ and $\mathbf{v}_j$. The cosine similarity then gives the embedding-based similarity: $s^{\text{LMEnc}}(o_i, p_j) = \cos(\mathbf{v}_i, \mathbf{v}_j)$. The final semantic similarity features used for fusion are: $\left\{s^{\text{LLM}}(o_i, p_j),\ s^{\text{LMEnc}}(o_i, p_j)\right\}$. These signals complement spatial and visual information, capturing subtle semantic variations such as brand abbreviations, semantics, and descriptive modifiers that frequently occur in real-world POI data.

This two-stage design, consisting of high-level filtering via SOTA LLMs and dense similarity scoring via language model encoders, enables both efficient pruning and semantic sensitivity. It helps preserve ambiguous yet potentially significant pairs for downstream classification, ultimately supporting robust ghost kitchen identification through alignment-based reasoning. In our implementation, we instantiate f_{LMEnc} using a pre-trained BERT model and apply it to both name and address fields.

D. Visual Consistency via Vision-Language Models

The final modality in LLM4GKID evaluates visual consistency between POI pairs using VLMs. This step leverages visual signals such as storefront photos, brand logos, or representative dishes to identify latent correspondences not captured by spatial or textual information. Due to the relatively high computational cost of VLM inference, visual matching is applied only to candidate pairs $\mathcal{C}'$ that have passed prior spatial and semantic filtering.

Let x^{vis} denote the visual content associated with a POI x, which may include one or more images. For each pair $(o_i, p_j) \in \mathcal{C}'$, we define a vision-language alignment function:

$$(v_{\text{bin}}, e_{\text{code}}) = f_{\text{VLM}}(o_i^{\text{vis}}, p_j^{\text{vis}}), \tag{5}$$

where $v_{\text{bin}} \in \{0, 1\}$ denotes the binary decision on visual consistency, and e_{code} encodes any exceptional condition such as missing, corrupted, or uninformative images. We adopt this binary formulation to prioritize reliability and robustness: during development we found that continuous confidence estimates from VLMs exhibit high variance across heterogeneous image types (e.g., storefronts vs. food photography), whereas binary decisions remain more stable and easier for downstream models to integrate.

To accommodate variation in visual content across platforms, we employ prompt-based comparison strategies that encourage the VLM to focus on brand-identifying features even when image styles differ. This enables consistent interpretation across mismatched visual modalities and improves alignment in cases where one platform emphasizes storefront exteriors and the other highlights dishes or interior scenes.

The resulting binary visual indicator is incorporated directly into the multimodal fusion stage, complementing spatial proximity and textual semantics. It is particularly effective in resolving borderline or conflicting cases where visual branding or iconography provides decisive evidence of identity. Our design remains fully modular: any VLM capable of structured pairwise comparison can serve as the instantiation of f_{VLM} within the proposed framework.

E. Supervised Fusion and Final Prediction

In the final stage of LLM4GKID, we integrate multimodal features via supervised learning to determine whether an online-offline POI pair $(o_i, p_j) \in \mathcal{C}'$ refers to the same real-world entity. For each candidate pair, we extract a set of alignment features encompassing spatial, semantic, and visual dimensions:

$$\mathbf{x}_{ij} = \big[\underbrace{h(o_i, p_j)}_{\text{spatial}}, \underbrace{s^{\text{VLM}}(o_i, p_j)}_{\text{visual}}, \underbrace{s^{\text{LLM}}(o_i, p_j),\ s^{\text{LMEnc}}(o_i, p_j)}_{\text{semantic}} \big], \tag{6}$$

where $h(o_i, p_j)$ denotes the haversine distance, s^{LLM} and s^{LMEnc} are semantic similarity scores, and s^{VLM} is the visual consistency score.

Let $f_{\text{sup},\phi} : \mathbf{x}_{ij} \to \{0, 1\}$ denote the supervised classifier parameterized by ϕ, which outputs a binary prediction indicating whether the pair corresponds to the same restaurant. The model is trained on a manually labeled dataset $\mathcal{D} = \{(\mathbf{x}_{ij}, y_{ij})\}$, where $y_{ij} \in \{0, 1\}$ is the ground-truth label. The training objective is to minimize the empirical classification loss:

$$\min_{\phi}\ \mathcal{L}_{\text{sup}}(\phi) = \sum_{(\mathbf{x}_{ij}, y_{ij}) \in \mathcal{D}} \ell\big(f_{\text{sup},\phi}(\mathbf{x}_{ij}),\ y_{ij}\big), \tag{7}$$

where $\ell(\cdot, \cdot)$ is a binary loss function such as cross-entropy.

At inference time, the classifier outputs predicted labels $\hat{y}_{ij} = f_{\text{sup}}(\mathbf{x}_{ij})$ for all pairs in $\mathcal{C}'$. Ghost kitchen detection then follows the alignment failure principle: for a given online POI $o_i \in \mathcal{O}$, if none of its candidate pairs is predicted as a match, i.e.,

$$f_{\text{sup}}(\mathbf{x}_{ij}) = 0 \quad \forall p_j \in \mathcal{C}'_i, \tag{8}$$

Then o_i is classified as a ghost kitchen.

This negative matching logic interprets consistent alignment failure as evidence of physical nonexistence. By unifying spatial, semantic, and visual cues within a supervised decision model, LLM4GKID enables scalable and interpretable detection of ghost kitchens across urban food ecosystems.

F. Learning Pipeline

Algorithm 1 outlines the complete LLM4GKID pipeline for identifying ghost kitchens via multimodal POI alignment. The process begins with spatial filtering based on a distance threshold τ_s, producing an initial candidate set $\mathcal{C}$ of geographically plausible POI pairs. A semantic filtering step follows, where an LLM-based name similarity score is computed for each pair.

Algorithm 1 LLM4GKID for Ghost Kitchen Identification

Require: Online POIs $\mathcal{O}$, Offline POIs $\mathcal{P}$, spatial and semantic thresholds τ_s, τ_m, LLM f_{LLM}, LM Encoder f_{LMEnc}, VLM f_{VLM}, classifier f_{sup}
Ensure: Matched pairs $\mathcal{A}$, Ghost kitchens $\mathcal{G}$

$\mathcal{C} \leftarrow \emptyset$ ▷ spatial
for all $o_i \in \mathcal{O}$ **do**
  for all $p_j \in \mathcal{P}$ **do**
    if $h(o_i^{\text{geo}}, p_j^{\text{geo}}) \le \tau_s$ **then**
      $\mathcal{C} \leftarrow \mathcal{C} \cup \{(o_i, p_j)\}$
    end if
  end for
end for
$\mathcal{C}' \leftarrow \{(o_i, p_j) \in \mathcal{C} \mid s^{\text{LLM}}(o_i, p_j) > \tau_m\}$ ▷ semantic
$\mathcal{A} \leftarrow \emptyset$
for all $(o_i, p_j) \in \mathcal{C}'$ **do**
  $s^{\text{LLM}}(o_i, p_j),\ s^{\text{LMEnc}}(o_i, p_j) \leftarrow f_{\text{LLM,LMEnc}}(o_i, p_j)$
  ▷ semantic
  $s^{\text{VLM}}(o_i, p_j) \leftarrow f_{\text{VLM}}(o_i, p_j)$ ▷ visual
  $\mathbf{x}_{ij} = \big[h(o_i, p_j),\ s^{\text{LLM}}(o_i, p_j),\ s^{\text{LMEnc}}(o_i, p_j),\ s^{\text{VLM}}(o_i, p_j) \big]$
  Train f_{sup} on $(\mathbf{x}_{ij}, y_{ij})$ ▷ classification
  if $f_{\text{sup}}(\mathbf{x}_{ij}) = 1$ **then**
    $\mathcal{A} \leftarrow \mathcal{A} \cup \{(o_i, p_j)\}$
  end if
end for
$\mathcal{G} \leftarrow \{o_i \in \mathcal{O} \mid \nexists (o_i, p_j) \in \mathcal{A}\}$ ▷ identify ghost kitchens
return $\mathcal{A}, \mathcal{G}$

Only pairs exceeding a semantic threshold τ_m are retained in the refined set $\mathcal{C}'$.

For each pair in $\mathcal{C}'$, we compute a full feature vector $\mathbf{x}_{ij}$ incorporating spatial distance, semantic similarity scores (from both LLM and encoder), and visual consistency scores from a vision-language model. These feature vectors are then passed to the supervised classifier f_{sup}, which produces binary predictions indicating whether each pair refers to the same underlying restaurant entity.

Aligned pairs are collected into the set $\mathcal{A}$, and ghost kitchens are inferred via the alignment failure principle: an online POI $o_i \in \mathcal{O}$ is labeled as a ghost kitchen if no match is found in $\mathcal{A}$. The output of the pipeline consists of the final alignment set $\mathcal{A}$ and the detected ghost kitchen set $\mathcal{G}$.

This modular, stepwise pipeline ensures both computational efficiency by progressively narrowing the candidate space and robustness by integrating heterogeneous signals under a unified supervised prediction framework.

IV. Experiments

A. Dataset Description

We derived restaurant POIs in Shenzhen from two Online Delivery Platforms (ODP) Meituan (MT, n=63,582) and Ele.me (ELE, n=45,076), which together hold over 95% of China's delivery market share—and from Dianping (DP,

n=174,445) as a Crowdsourced Review Platform (CRP), during September 1–7, 2024. After excluding dessert, snack, and beverage-only outlets and filtering out entries lacking dine-in verification, we retained 141,859 physically verified establishments. Each POI includes spatial coordinates, semantic descriptors, and storefront images for multimodal analysis. We demonstrate LLM4GKID's matching on the MT–DP pair; identical pipelines were applied to ElE-DP and ELE–MT to detect additional ghost kitchens and remove duplicates, but detailed results focus on MT–DP to avoid redundancy.

B. Annotation Protocol and Quality Control

We generated approximately 2.1 million candidate ODP–CRP pairs using spatial filtering (150 m radius) and a k-nearest-neighbor search. To reduce computational overhead while preserving evaluation accuracy, we adopted an active and comprehensive sampling strategy designed to avoid bias from relying on a single signal. Importantly, to focus annotation effort on informative and ambiguous cases, we first prune trivial identical-name pairs from the manual annotation pool, as these pairs are overwhelmingly straightforward and contribute little to evaluating model robustness.

The final annotated dataset was constructed from three complementary sampling streams, described below: *(1) LLM-filtered high-similarity pairs.* The first sampling stream consists of pairs that exhibit relatively high textual similarity according to the LLM-based semantic model. These candidates commonly arise in cross-platform alignment and include both true matches and ambiguous non-matches that require careful verification. From this pool, a stratified random sample was selected for annotation. *(2) "Name-drift" true matches.* The second sampling stream targets true matches that refer to the same physical restaurant but have inconsistent or substantially different names across platforms. We identify these cases using phone-number alignment within 150 m and select pairs with low semantic similarity but shared phone numbers. Annotators manually verify each pair using storefront images and external map sources to ensure that these examples represent true naming inconsistencies rather than ghost kitchens. Including these cases ensures that the annotated dataset captures hard positive examples that would not be surfaced by semantic filtering alone. *(3) Human-in-the-loop active sampling.* The third sampling stream focuses on cases the initial classifier finds most uncertain. After training a preliminary model on the first two streams, we applied it to the full candidate pool and selected pairs whose predicted match probabilities fell near the decision boundary, specifically those with $0.4 \leq \hat{p} \leq 0.6$. These borderline cases often exhibit conflicting spatial, textual, or visual signals and represent examples that automated filtering tends to exclude. Annotators label these uncertainty-selected pairs following the same protocol, ensuring that the dataset includes a broad spectrum of challenging edge cases.

Annotation procedure. Pairs from all sampling streams were labeled by four trained annotators using a web-based interface displaying map locations, textual attributes, and storefront images. Labels included *Match*, *Non-match*, and *Unsure*. This process yielded 3,994 high-quality ground-truth pairs with an approximately balanced class ratio (*Match:Non-match* $\approx 3:2$). Disagreements were resolved via majority voting, while *Unsure* cases were escalated for group discussion. For all *Non-match* labels, annotators performed an additional verification step by inspecting the top twenty spatial candidates and conducting targeted online searches to avoid mislabeling due to incomplete information.

Generalization test set. To evaluate robustness beyond the primary city–time–platform setting, we construct a additional held-out test sets targeting cross-city, temporal, and cross-platform generalization. The test set contains 593 annotated pairs (above 10% of the main dataset) and is randomly selected from the candidate pairs pool (C') of new city (Shanghai) in a earlier temporal snapshot (Oct 2023), and an alternative platform (ELE). These sets preserve a consistent mix of easy and difficult cases while keeping annotation costs tractable. They are not used for training and are evaluated separately in the Section Robustness Analysis.

Finally, we performed an 80%–10%–10% stratified split into training, validation, and test sets while preserving class balance and ensuring proportional representation from all sampling streams.

C. Baseline Methods and Evaluation Framework

Baseline Method Selection. To comprehensively evaluate LLM4GKID's effectiveness, we compare against three state-of-the-art methods representing different POI conflation paradigms, adapted for our ghost kitchen detection task. The POI Data Fusion method by Wang et al. [22] represents traditional machine learning approaches that combine spatial and non-spatial features through Random Forest weighting, focusing on feature engineering and ensemble learning. ESRM by Li et al. [29] exemplifies transformer-based entity matching; we adapt its deep semantic understanding approach through pretraining and fine-tuning strategies to our POI conflation context. PlacERN by Cousseau & Barbosa [27] represents multi-view deep learning approaches; we adapt its specialized neural encoders for integrating multiple data modalities to our restaurant-specific matching task, balancing computational efficiency with representation richness. The detailed comparison of modalities, methodologies, components, and performance characteristics is presented in Table I.

Evaluation Metrics and Protocol. Our evaluation framework examines performance across multiple dimensions. We report precision, recall, F1-score, and cross-validated F1 (CV F1) for all baseline models, and apply 5-fold cross-validation on the training set for hyperparameter optimization. To assess multimodal integration within LLM4GKID, we evaluate the behavior of several fusion classifiers under stepwise modality addition, including Logistic Regression, Random Forest, LightGBM, and XGBoost. These classifiers differ in their capacity to capture nonlinear interactions, handle high-dimensional inputs, and integrate heterogeneous feature types. We further conduct a cost-effectiveness analysis to characterize scalability and to examine the relationship among cost, runtime, and predictive accuracy for potential deployment preferences. Ablation and interpretability analyses quantify the

TABLE I: Comparison of Models Adapted for POI Conflation

Model	Modalities	Methodology	Key Components	Performance
POI Data Fusion [22]	Spatial (coords) Non-spatial (name, address)	Multi-feature similarity Random Forest	Hybrid name similarity Jaccard on address Euclidean distance RF weight learning	Higher accuracy & recall vs. baselines
ESRM [29]	Textual (name, address, category)	Transformer pretrain Fine-tune	RCP pretraining RLM pretraining Attribute alignment	Category F1 $>$ 90% Address BLEU $>$ 95%
PlacERN [27]	Text (name, address) Category labels Spatial (haversine)	Multi-view deep encoders	BiGRU encoder CNN encoder Category embedding Geo embedding Feed-forward classifier	Outperforms RF/LightGBM on $F_{0.5}$, Gini, AUPR
LLM4GKID (This work)	Spatial (haversine) Semantic (LLM + BERT) Visual (VLM)	Stepwise multimodal filtering Supervised fusion	1. Spatial blocking 2. LLM filtering 3. LM encoder 4. VLM consistency 5. Supervised fusion	See evaluation in subsection IV-E

marginal contribution of each modality to overall performance. Robustness is assessed using generalization results on the held-out test set and a spatial consistency analysis of low-confidence matches, ensuring that the conservative filtering strategy identifies potential ghost kitchens without introducing spatial bias.

D. Implementation and Experimental Setup

Parameter Settings. We use GPT-4.1 Mini with custom prompt templates (as shown in Appendix A & B) for semantic and visual matching, as it provides a practical and stable setup for large-scale batch inference on the online platform without deployment requirements. A broader comparison of alternative LLM–VLM configurations is presented in the Quantitative Performance Comparison. To ensure robust performance, we employed grid search with five-fold cross-validation, optimizing average F1-score across POI conflation and ghost detection. The detailed model hyperparameters are specified in Appendix C. We fixed the random seed to 42 for reproducibility and selected the highest-performing configuration for final evaluation.

E. Quantitative Performance Comparison

Baselines Comparison. We begin by benchmarking our approach against three adapted state-of-the-art baselines using held-out test data. Figure 2 summarizes each model's F1-score, precision, recall, and cross-validated (CV) F1-score: 1) POI Data Fusion (F1 = 0.674) relies on hand-tuned spatial/name/address heuristics and Random Forest weighting. While this approach demonstrates solid foundational principles for multi-feature integration, its performance may be affected by the inherent name similarity of the nearby restaurant, given that more than half the feature importance (0.558) of name features is detected in their original model. 2) The PlacERN method (F1 = 0.698) represents a sophisticated deep learning architecture specifically designed for place deduplication, utilizing multi-view encoders that learn distinct representations from different information levels. Their approach

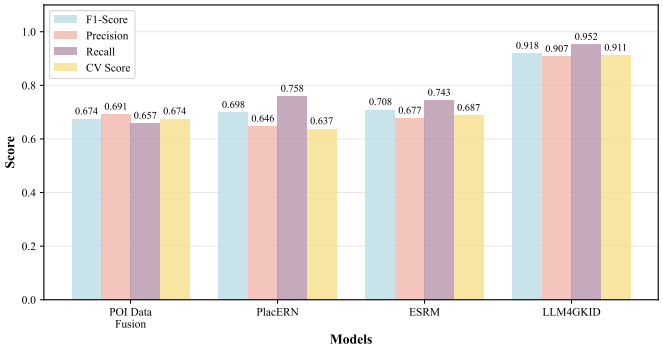


Fig. 2: Performance Comparison between LLM4GKID Approach & Baseline Methods

demonstrated remarkable effectiveness in handling challenging datasets with missing values and high class imbalance, consistently outperforming competitive algorithms across multiple evaluation metrics. The modest 3.6 percentage points (pp) improvement over POI Data Fusion suggests that their multi-view encoding strategy faces challenges in our restaurant-focused onctext. PlacERN's architecture was particularly optimized for general place deduplication across diverse POI categories, where categorical distinctions provide strong discriminative signals. Our restaurant-focused dataset presents a more constrained semantic space, where establishments share similar categorical properties, potentially limiting the effectiveness of their multi-view approach, which excels when processing heterogeneous place types with distinct characteristics. 3) ESRM (F1 = 0.708) applies a sophisticated transformer-based textual alignment paradigm and achieves balanced performance with precision (0.677) and recall (0.743). The 1.0 pp improvement over PlacERN demonstrates the value of advanced textual representations, though the model still faces challenges in the specialized context of restaurant disambiguation where subtle semantic distinctions are critical. 4) LLM4GKID (F1 = 0.918) builds upon these established methodologies and incorporates LLM-driven semantic similarity and visual con-

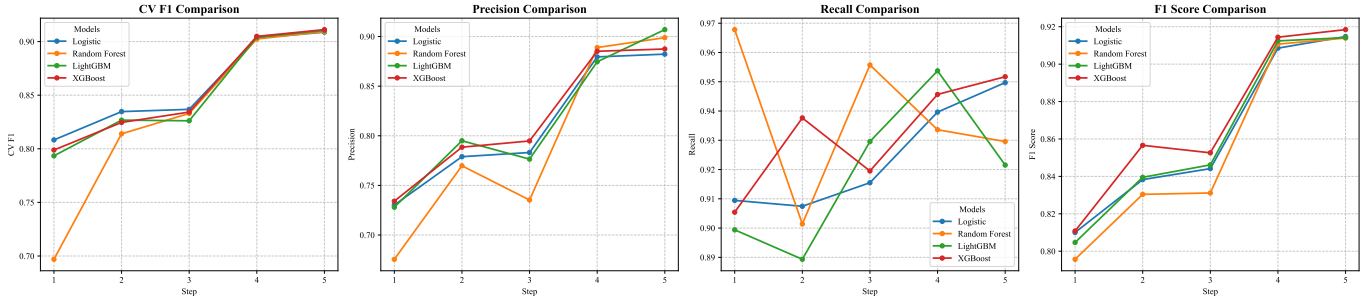


Fig. 3: Supervised Fusion Models Comparison of the LLM4GKID Approach

sistency to address the specific challenges of restaurant-only matching. The substantial 21.0 pp improvement over ESRM, combined with high precision (0.907) and exceptional recall (0.952), demonstrates that multimodal fusion provides complementary information when traditional categorical and spatial distinctions become insufficient for disambiguation within homogeneous establishment types. The strong cross-validation score (CV F1 = 0.911) further confirms the model's robust generalization strength. Overall, the baseline methods exhibit lower performance than reported in their original tasks, primarily because our dataset is constructed using hard example mining and includes highly confusable restaurant pairs without category attributes, making the matching task substantially more challenging.

Fusion Classifier Comparison. We benchmarked four supervised fusion models, including Logistic Regression, Random Forest, LightGBM, and XGBoost across five fusion stages (Figure 3): Geo Distance (Step 1), Name Similarity (BERT) (Step 2), Address Similarity (BERT) (Step 3), Name Similarity (LLM) (Step 4), and Visual Consistency (VLM) (Step 5). Across the first three stages, all models achieve broadly comparable performance, with CV F1 scores concentrated in the 0.79–0.84 range. Geo Distance alone yields CV F1 scores around 0.70–0.81, and adding BERT-based name and address similarity steadily lifts all models to 0.83 by Step 3. Among them, XGBoost exhibits consistently strong precision, while Random Forest achieves the highest recall in the early stages. The introduction of LLM-based name similarity at Step 4 produces the most significant performance jump. All four models benefit, but XGBoost and LightGBM show the largest gains: XGBoost's CV F1 increases from 0.834 (Step 3) to 0.905. Logistic Regression also improves (CV F1 = 0.903), but remains slightly behind the boosted trees. Overall, the results show that while Logistic Regression remains a competitive and lightweight baseline, gradient-boosted tree models, particularly XGBoost, most effectively exploit the combined spatial, textual, and visual signals, achieving the highest CV F1 in the final fusion stage.Thus, XGBoost were used as the default fusion model for the further analysis.

Cost-effectiveness Analysis. To evaluate the practical deployment efficiency of LLM4GKID, we analyse the joint relationship between inference cost, runtime, and predictive performance using the model configurations shown in Figure 4. Across the full grid of open-source and closed-source LLMs

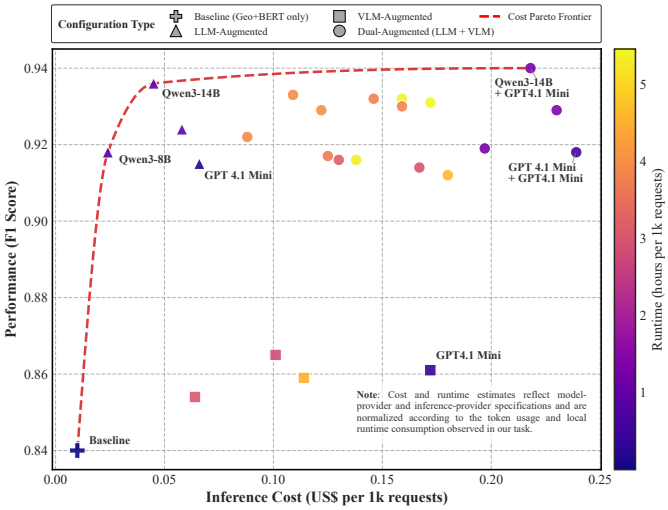


Fig. 4: Cost Effectiveness Analysis of LLMs and VLMs

(Qwen3-8B/14B/32B and GPT-4.1 Mini) paired with VLMs (Qwen3-VL-4B/8B/32B and GPT-4.1 Mini), the fusion model (XGBoost) results exhibit a clear Pareto frontier that highlights the diminishing marginal returns of increasingly expensive models. While the baseline Geo (i.e., spatial distance)+BERT provides only modest accuracy at minimal cost, the most cost-effective gains come from *LLM-augmented* variants such as Qwen3-8B and Qwen3-14B, which deliver substantial performance improvements (F1: 0.90–0.94) at comparatively modest increases in inference cost. By contrast, *VLM-only* augmentation yields smaller incremental gains with substantially higher compute requirements, reflecting the heavier runtime footprint of visual–language models relative to text-only LLMs. One of the most computationally intensive *dual-augmented* combinations (e.g., Qwen3-14B + GPT-4.1 Mini VLM) achieve the highest absolute performance but sit at the upper-right of the efficiency frontier [39], making them appropriate primarily for scenarios where maximum accuracy outweighs real-time constraints.

Overall, the frontier analysis shows that LLM4GKID achieves near-optimal performance at significantly lower cost by (i) aggressively reducing candidate pairs through spatial filtering and multi-stage semantic pruning, and (ii) applying expensive VLM inference only to a small, high-uncertainty subset of pairs. This design ensures both accuracy and op-

erational scalability, with an estimated city-scale deployment (for $\approx 140K$ inference requests) cost of approximately US$7 inference cost for the most cost-effective configuration (Baseline + Qwen3-14B for semantic task) and US$33 for the highest-accuracy configuration (Baseline + Qwen3-14B for the semantic & GPT4.1 Mini for visual task) in Shenzhen.

F. Ablation and Interpretability Analysis

To systematically assess the relative contribution of each modality to LLM4GKID's overall performance, we conducted a comprehensive leave-one-out ablation study on the full model configuration (spatial + semantic (BERT) + semantic (LLM) + visual) using XGBoost as the fusion model. The results, summarized in Table II, reveal a clear hierarchical importance of different modalities in our progressive framework.

TABLE II: Ablation results: F1 and AUC when removing one modality at a time.

Model Variants	F1 Score	AUC	ΔF1 (vs. full)
LLM4GKID	0.9184	0.9522	—
– w/o spatial	0.9084	0.9417	–0.0101
– w/o semantic (BERT)	0.9106	0.9368	–0.0078
– w/o semantic (LLM)	0.8684	0.8974	–0.0500
– w/o visual	0.9144	0.9473	–0.0040

Ablation Findings. The most striking observation is the dominant role of LLM-based semantic similarity. Removing this feature leads to the largest drop in performance (F1: –0.0500), substantially greater than the impact caused by removing spatial, BERT-based semantic, or visual features. This confirms that advanced LLM reasoning is essential for resolving nuanced linguistic variations, such as synonyms, dialectal differences, and menu-driven descriptions, that commonly arise between ODP and CRP restaurant names.

Spatial proximity emerges as the second most crucial modality, with its exclusion yielding a substantial F1 decrease (-0.0101). This confirms that geographic filtering not only removes implausible candidates but also provides decisive matching signals, ensuring that higher-cost semantic and visual processing stages operate only within a meaningful search scope. BERT-based semantic features and VLM visual consistency each contribute modest but measurable improvements (F1 reductions of –0.0078 and –0.0040 when removed, respectively). Their smaller contributions highlight two challenges: (1) BERT, as a traditional language model, struggles to distinguish subtle naming variations in dense urban environments where nearby restaurants frequently share similar naming conventions (Tobler's law). (2)Visual cues, while informative, are inherently sparse, occasionally inconsistent in quality, and computationally expensive, which limits their marginal gains after strong spatial and LLM filtering. Overall, the ablation study shows that LLM semantics and spatial filtering form the backbone of the LLM4GKID pipeline, while BERT and VLM features act as important, but secondary refinements.

Feature Interpretability Analysis. To further understand how the model internally leverages these modalities, we conduct a comprehensive SHAP-based interpretability analysis, as shown in Figure 5. The results closely mirror the ablation findings. Among all features, LLM-based name similarity exhibits the highest global SHAP importance, with a sharp, monotonic increase in positive contribution once similarity exceeds approximately 65%. This pattern directly corresponds to the substantial performance drop observed when removing LLM semantics in the ablation study, reinforcing its decisive role in distinguishing matched from unmatched restaurant pairs. Geo-distance emerges as the second most influential factor, with SHAP values declining rapidly beyond approximately 20 meters, confirming its function as a strong early-stage filter that suppresses unlikely candidates. BERT-based name similarity and address similarity contribute at moderate levels, offering valuable refinements in cases where LLM signals alone remain ambiguous. In contrast, visual consistency (VLM) exhibits lower but still interpretable SHAP contributions, primarily enhancing borderline predictions when both naming and address cues are insufficient. The SHAP interaction heatmap further shows that cross-feature interactions are generally weak, suggesting that each modality contributes largely independently. A mild interaction (0.071) between geographical distance (GD) and BERT-based name similarity (NSB), together with the observation that high NSB values sometimes correspond to relatively low SHAP contributions, reflects the challenge posed by densely clustered commercial areas, where neighbouring establishments frequently share highly similar names . Together, these interpretability patterns corroborate the hierarchical importance revealed through ablation and demonstrate that the model's decision process aligns closely with the intended progressive, multi-modal design of LLM4GKID.

G. Robustness Analysis

To evaluate whether LLM4GKID remains reliable outside the primary training conditions, we conduct two complementary robustness analyses. The first examines generalization performance under domain shifts in geography, time, and platform source. The second investigates spatial robustness by testing whether uncertain or potentially misclassified POIs exhibit spatial clustering that could bias downstream ghost-kitchen location analysis. Together, these analyses provide a comprehensive assessment of the model's stability across heterogeneous urban contexts and spatial distributions.

(1) Generalization. We first assess out-of-domain generalization using a held-out evaluation set constructed to probe variation across city, time, and platform. This set consists of 587 manually annotated POI pairs sampled from Shanghai, drawn from an earlier temporal snapshot (October 2023), and supplemented with cases from an alternative ODP (Ele.me). These data were never used in training and reflect realistic domain shifts in naming conventions, storefront styles, and platform-specific visual patterns.

The model exhibits strong generalization under these conditions, achieving an F1 of 0.9193, AUC of 0.9729, precision of 0.9934, and a recall of 0.8555. Importantly, the performance on this held-out set is comparable to, and in some aspects higher than, the results observed in Shenzhen. This difference is expected because the Shenzhen dataset was constructed

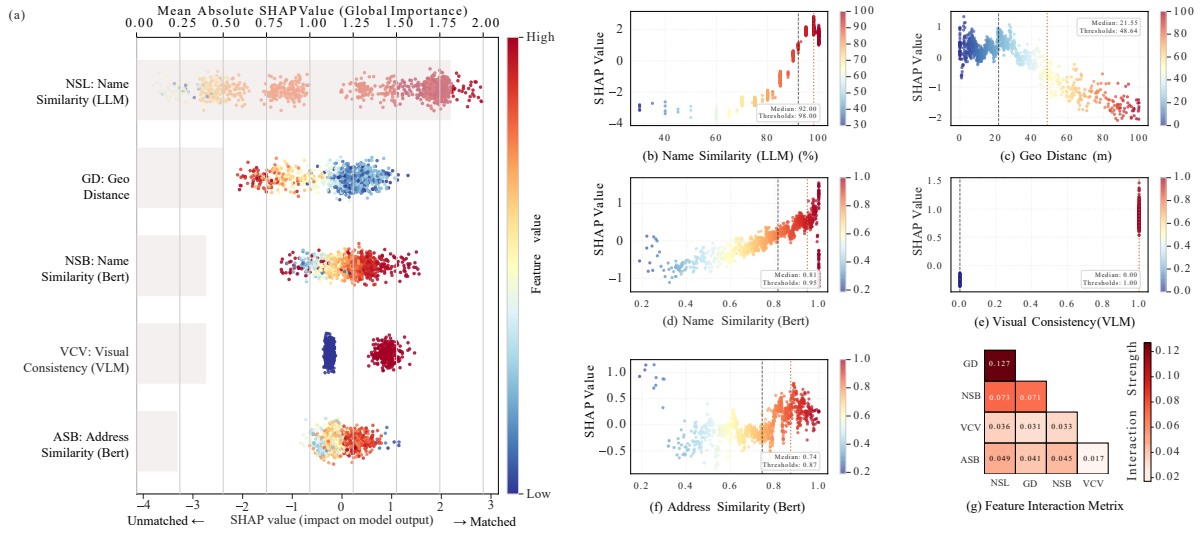


Fig. 5: The Feature Interpretability Analysis of the LLM4GKID Approach

using an active sampling strategy that deliberately concentrates difficult borderline cases, while the Shanghai set reflects a more natural distribution of easy and moderately challenging examples. The model's high performance in this out-of-domain setting demonstrates that the multimodal fusion framework generalizes effectively beyond the original training environment.

(2) Spatial Robustness. Spatial bias in classification errors could compromise the validity of ghost-kitchen location analysis, as algorithmic failures might create artificial geographic clustering patterns unrelated to actual business operations. To verify that low-confidence POIs do not cluster spatially, we examined spatial autocorrelation on the set of online POIs whose best candidate match had a predicted probability in the uncertainty interval ($0.4 \leq \max_{p_j \in \mathcal{C}'_i} \hat{p}_{ij} \leq 0.6$) (typical cases as shown in Appendix D). We calculated Moran's I [40] on the uncertain caeses using a distance threshold of 500 m. The resulting Moran's I = 4.8×10^{-5} ($z = 0.03, p = 0.97$) indicates no significant global autocorrelation of predictive probability (i.e., errors are spatially random). The negligible Moran's I demonstrates that LLM4GKID's misalignments do not exhibit spatial dependency. This spatial randomness confirms that the aligned dataset can be used reliably for downstream spatial analyses without introducing method-driven clustering artifacts.

H. Discussion

Our experiments reveal three key insights about multimodal POI alignment and ghost-kitchen identification that advance current knowledge while establishing explicit connections to recent entity-matching research.

Multimodal signals are decisive. LLM4GKID attains an F1 score of 0.918, surpassing every single-modality ablation. The majority of improvement stems from LLM-generated semantic signals. This echoes Cheng et al.[35]'s finding that deep language understanding enriches POI representations, extending this insight to expose establishments with divergent digital and physical footprints. By applying LLM (e.g., Qwen3-8B or 14B)'s capabilities to the hard example candidate set after spatial filtering, we capture nuanced reasoning for untangling colloquial restaurant names and brand aliases, yielding a twenty-point F1 margin over classical matchers without prohibitive computational overhead.

Progressive filtering enables metropolitan-scale analytics. We shrink the candidate search space from roughly fifteen billion raw pairs to 2.1 million with spatial blocking and then to about two hundred thousand with conservative language-model filtering, a 90% cut relative to the already standard blocking stage. This staged pruning strategy maintains a recall above 0.95 while significantly reducing inference cost, enabling continuous monitoring of large urban areas.

Ethical and Practical Risks The application of the framework needs careful ethical and operational considerations. Data collection must comply with platform terms of service and local regulations governing digital data use. Although human-in-the-loop annotation improves data quality and model robustness, it also introduces subjective judgment, and mislabeling may propagate harmful biases into downstream analyses. Ensuring annotator training and transparent labeling protocols is therefore essential. In addition, automatic matching systems risk incorrectly identifying legitimate businesses as ghost kitchens, which may have reputational or regulatory consequences. Responsible deployment should incorporate conservative thresholds, human review of ambiguous cases, and clear communication of uncertainty with collaborative and integrated governance in real-world practices [41, 42].

V. Conclusion

This study introduces LLM4GKID, a multimodal framework for detecting ghost kitchens through staged fusion of spatial, semantic, and visual signals. By integrating geographic filtering, language model–based similarity assessment, and VLM-driven visual consistency, the approach reliably identifies delivery-only establishments that lack verified counterparts in review platforms. Results on manually annotated

Shenzhen data show strong improvements over existing POI conflation methods, with efficient candidate filtering, strong generalization strength and no detectable spatial bias in misclassifications. Although the framework remains constrained by platform-specific data availability, limited examples of ambiguous edge cases, it demonstrates a robust and scalable solution for uncovering establishments with asymmetric digital footprints, with potential applications to food-access analysis, regulatory monitoring, and other digital–physical mismatches such as dark stores or virtual offices.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (52408095), the Hong Kong Research Grants Council (17601425; 27601324), Guangdong Basic and Applied Basic Research Foundation (2025A1515010902), and the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (2022ONRC001).

References

[1] R. Cai, X. Y. Leung, and C. G.-Q. Chi, "Ghost kitchens on the rise: Effects of knowledge and perceived benefit-risk on customers' behavioral intentions," *International Journal of Hospitality Management*, vol. 101, p. 103110, 2022. 1, 2, 4

[2] A. Shapiro, "Platform urbanism in a pandemic: Dark stores, ghost kitchens, and the logistical-urban frontier," *Journal of Consumer Culture*, vol. 23, no. 1, pp. 168–187, Feb. 2023. 1

[3] Z. Laheri, I. Ferris, D. T. da Cunha, and J. M. Soon-Sinclair, "'going dark'or under the radar? challenges and opportunities for local authorities and dark kitchens in ensuring food safety," *Food Control*, p. 111179, 2025. 1

[4] X. Dai and L. Wu, "The impact of capitalist profit-seeking behavior by online food delivery platforms on food safety risks and government regulation strategies," *Humanities and Social Sciences Communications*, vol. 10, no. 1, pp. 1–12, 2023. 1, 2

[5] O. T. K. Vu, A. D. Alonso, T. D. Tran, and G. J. Nicholson, "Illuminating the dark kitchen business model-a knowledge-based perspective from the supply-side," *Journal of Hospitality and Tourism Management*, vol. 55, pp. 318–331, 2023. 1

[6] M. P. Hakim, V. M. D. Libera, L. D. Zanetta, E. Stedefeldt, L. M. Zanin, J. M. Soon-Sinclair, M. Z. Wiśniewska, and D. T. Da Cunha, "Exploring dark kitchens in brazilian urban centres: A study of delivery-only restaurants with food delivery apps," *Food Research International*, vol. 170, p. 112969, 2023. 1, 2

[7] Y. Kim, M. Lee, B.-D. Kim, and T. Roh, "Power of agglomeration on electronic word–of–mouth in the restaurant industry: Exploring the moderation role of review quality difference," *Journal of Retailing and Consumer Services*, vol. 78, p. 103759, 2024. 1

[8] Y. Tang, Z. Wang, A. Qu, Y. Yan, Z. Wu, D. Zhuang, J. Kai, K. Hou, X. Guo, J. Zhao *et al.*, "Itinera: Integrating spatial optimization with large language models for open-domain urban itinerary planning," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, 2024, pp. 1413–1432. 1, 3

[9] L. He, H. Li, and R. Zhang, "A semantic-spatial aware data conflation approach for place knowledge graphs," *ISPRS International Journal of Geo-Information*, vol. 13, no. 4, p. 106, 2024. 1

[10] A. Psyllidis, S. Gao, Y. Hu, E.-K. Kim, G. McKenzie, R. Purves, M. Yuan, and C. Andris, "Points of interest (poi): A commentary on the state of the art, challenges, and prospects for the future," *Computational urban science*, vol. 2, no. 1, p. 20, 2022. 1, 2

[11] T. L. Lei, "Large scale geospatial data conflation: A feature matching framework based on optimization and divide-and-conquer," *Computers, Environment and Urban Systems*, vol. 87, p. 101618, 2021. 1

[12] C. Ling, X. Niu, J. Yang, J. Zhou, and T. Yang, "Unravelling heterogeneity and dynamics of commuting efficiency: Industry-level insights into evolving efficiency gaps based on a disaggregated excess-commuting framework," *Journal of Transport Geography*, vol. 115, p. 103820, 2024. 1

[13] D. O. Hassan and B. A. Hassan, "A comprehensive systematic review of machine learning in the retail industry: classifications, limitations, opportunities, and challenges," *Neural Computing and Applications*, vol. 37, no. 4, pp. 2035–2070, 2025. 2

[14] K. Sun, Y. Hu, Y. Ma, R. Z. Zhou, and Y. Zhu, "Conflating point of interest (poi) data: A systematic review of matching methods," *Computers, Environment and Urban Systems*, vol. 103, p. 101977, 2023. 2, 3

[15] Z. Huang, "Disambiguate entity matching using large language models through relation discovery," in *Proceedings of the Conference on Governance, Understanding and Integration of Data for Effective and Responsible AI*, 2024, pp. 36–39. 2

[16] L. Nield, H. Martin, C. Wall, J. Pearce, R. Rundle, S. Bowles, D. Harness, and J. D. Beaumont, "Consumer knowledge of and engagement with traditional takeaway and dark kitchen food outlets," *NIHR Open Research*, vol. 4, p. 64, 2025. 2, 4

[17] D. Kim, L. Dolega, and J. Park, "Dark kitchens and streetscapes: Exploring the location choices of 'dark kitchens' using street view imagery," *Applied Geography*, vol. 185, p. 103805, 2025. 2

[18] Y. Huang, T. R. Bishop, J. Adams, S. Cummins, M. Keeble, C. Rinaldi, A. Schiff, and T. Burgoine, "Understanding the socio-spatial distribution of "dark retail" in england: Development of a unique retail location dataset," *Health & place*, vol. 94, p. 103462, 2025. 2

[19] A. Ennis, L. Chen, C. D. Nugent, G. Ioannidis, and A. Stan, "High-level geospatial information discovery and fusion for geocoded multimedia," *International Journal of Pervasive Computing and Communications*, vol. 9, no. 4, pp. 367–382, 2013. 3

[20] N. Barret, F. Duchateau, F. Favetta, and L. Moncla, "Spatial entity matching with geoalign (demo paper)," in *Proceedings of the 27th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, 2019, pp. 580–583. 3

[21] N. Wang, J. Zeng, M. Chen, and S. Zhu, "An efficient algorithm for spatio-textual location matching," *Distributed and Parallel Databases*, vol. 38, no. 3, pp. 649–666, 2020. 3

[22] Y. Wang, C. Li, H. Zhang, B. Guo, X. Wei, and H. Zhao, "Poi data fusion method based on multi-feature matching and optimization," *ISPRS International Journal of Geo-Information*, vol. 14, no. 1, p. 26, 2025. 3, 7, 8

[23] R. Low, Z. D. Tekler, and L. Cheah, "An end-to-end point of interest (poi) conflation framework," *ISPRS International Journal of Geo-Information*, vol. 10, no. 11, p. 779, 2021. 3

[24] G. McKenzie, K. Janowicz, and B. Adams, "A weighted multi-attribute method for matching user-generated points of interest," *Cartography and Geographic Information Science*, vol. 41, no. 2, pp. 125–137, 2014. 3

[25] X. Xing, H. Lin, F. Zhao, and S. Qiang, "Local poi matching based on knn and lightgbm method," in *2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*. IEEE, 2022, pp. 455–458. 3

[26] Z. Li, N. Ding, C. Liang, S. Cao, M. Zhai, R. Huang, Z. Zhang, and B. Hu, "A semi-supervised framework fusing multiple information for knowledge graph entity alignment," *Expert Systems with Applications*, vol. 259, p. 125282, 2025. 3

[27] V. Cousseau and L. Barbosa, "Linking place records using multi-view encoders," *Neural Computing and Applications*, vol. 33, no. 18, pp. 12 103–12 119, 2021. 3, 5, 7, 8

[28] P. Li, J. Liu, A. Luo, Y. Wang, J. Zhu, and S. Xu, "Deep learning method for chinese multisource point of interest matching," *Computers, Environment and Urban Systems*, vol. 96, p. 101821, 2022. 3

[29] P. Li, Y. Wang, J. Liu, A. Luo, S. Xu, and Z. Zhang, "Enhanced semantic representation model for multisource point of interest attribute alignment," *Information Fusion*, vol. 98, p. 101852, 2023. 3, 7, 8

[30] Y. Tang, A. Qu, Z. Wang, D. Zhuang, Z. Wu, W. Ma, S. Wang, Y. Zheng, Z. Zhao, and J. Zhao, "Sparkle: Mastering basic spatial capabilities in vision language models elicits generalization to spatial reasoning," in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 2025, pp. 4083–4103. 3

[31] Y. Cao, J.-A. Yang, A. Nara, and M. M. Jankowska, "Designing and evaluating a hierarchical framework for matching food outlets across multi-sourced geospatial datasets: a case study of san diego county," *Journal of Urban Health*, vol. 101, no. 1, pp. 155–169, 2024. 3

[32] M. Trokhymovych and O. Kosovan, "Geodd: End-to-end spatial data de-duplication system," in *Proceedings of the Computational Methods in Systems and Software*. Springer, 2023, pp. 717–727. 3

[33] B. Xie, X. Ma, X. Shan, A. Beheshti, J. Yang, H. Fan, and J. Wu, "Multiknowledge and llm-inspired heterogeneous graph neural network for fake news detection," *IEEE Transactions on Computational Social Systems*, vol. 12, no. 2, pp. 682–694, 2025. 3

[34] Y. Chen, B. Chi, C. Li, Y. Zhang, C. Liao, X. Chen, and N. Xie, "Toward interactive next location prediction driven by large language models," *IEEE Transactions on Computational Social Systems*, pp. 1–17, 2025. 3

[35] J. Cheng, J. Wang, Y. Zhang, J. Ji, Y. Zhu, Z. Zhang, and X. Zhao, "Poi-enhancer: An llm-based semantic enhancement framework for poi representation learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 39, no. 11, 2025, pp. 11 509–11 517. 3, 11

[36] R. Peeters and C. Bizer, "Using chatgpt for entity matching," in *European Conference on Advances in Databases and Information Systems*. Springer, 2023, pp. 221–230. 3

[37] Z. Zhang, P. Groth, I. Calixto, and S. Schelter, "A deep dive into cross-dataset entity matching with large and small language models," in *Proceedings 28th International Conference on Extending Database Technology, EDBT 2025, Barcelona, Spain, March 25-28, 2025*, A. Simitsis, B. Kemme, A. Queralt, O. Romero, and P. Jovanovic, Eds. OpenProceedings.org, 2025, pp. 922–934. 3

[38] B. Berjawi, "Integration of heterogeneous data from multiple location-based services providers: A use case on tourist points of interest," Ph.D. dissertation, Université de Lyon, 2017. 5

[39] Y. Xu, C. Chen, W. Deng, L. Dai, and T. Yang, "Spatial eco-socio-economic trade-offs inform differentiated management strategies in mega-urban agglomerations," *npj Urban Sustainability*, vol. 5, no. 1, p. 43, 2025. 9

[40] Y. Chen, "Spatial autocorrelation equation based on moran's index," *Scientific Reports*, vol. 13, no. 1, p. 19296, 2023. 11

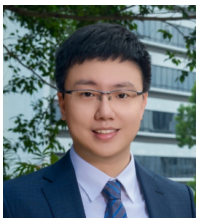
[41] C. Zhou, C. Richardson-Barlow, L. Fan, H. Cai, W. Zhang, and Z. Zhang, "Towards organic collaborative governance for a more sustainable environment: Evolutionary game analysis within the policy implementation of china's net-zero emissions goals," *Journal of Environmental Management*, vol. 373, p. 123765, 2025. 11

[42] J. Qu, T. Yang, K.-M. Nam, E. Kim, Y. Chen, and X. Liu, "Transport network changes and varying socioeconomic effects across china's yangtze river delta," *Journal of Transport Geography*, vol. 121, p. 104051, 2024. 11

Weipeng Deng earns his Bachelor's degree in Human Geography and Urban-rural planning from South China Normal University, and received his Master's degree in Urban Planning from the University of Melbourne. He is currently a PhD candidate in the Department of Urban Planning and Design at The University of Hong Kong and visiting PhD student in Urban Analytics Lab at National University of Singapore. He also serves as Decision Consulting Expert in Chinese Society for Urban Studies. His research explores urban sensing in the era of digital transformation, multimodal urban data fusion, GeoAI, and computational urban models.

Yihong Tang is a Ph.D. student in Transportation Engineering at McGill University. He earned his M.Phil. in Urban Analytics and Smart Cities from the University of Hong Kong and his B.Eng. in Computer Science and Technology from BUPT. He has held research positions at the Mobility AI Lab (HK PolyU), JTL Transit Lab at MIT. His research focuses on data-driven, statistical, and AI methods for designing connected, autonomous, and human-centered urban and transportation systems. His main interests include human-centered modeling (with equity, privacy, and intent awareness), large language models for urban science, multimodal transportation systems via data fusion, and spatiotemporal data modeling for forecasting, control, and behavior inference.

Chaofan Wang was born in Wenzhou, Zhejiang, China, in 1999. He received the bachelor's degree in landscape architecture from Zhejiang University, Hangzhou, Zhejiang, China, in 2021, and the master's degree in urban regeneration and development from the University of Manchester, Manchester, United Kingdom, in 2022. His major field of study is urban analytics. He is currently pursuing the Ph.D. degree in the Department of Urban Planning and Design, The University of Hong Kong, Hong Kong. His research interests include urban analytics and smart cities, human mobility, and activity-based modeling.



Tianren Yang (Member, IEEE) holds a Ph.D. in architecture (applied urban modelling) from the University of Cambridge. He earned an M.Sc. in urban design from the Georgia Institute of Technology, and completed both an M.Eng. in urban planning and a B.Eng. in landscape architecture at Tongji University.

He is currently an Assistant Professor and Assistant Head in the Department of Urban Planning and Design at the University of Hong Kong. His research centers on urban analytics and predictive modeling, with a particular focus on integrating behavioral data, urban technology, and planning policy to explain and anticipate urban dynamics. Through this work, he aims to develop generalizable frameworks and evidence-based tools that deepen understanding of how cities evolve and support more adaptive and effective planning practice.

Dr. Yang is a Chartered Member of the Royal Town Planning Institute and a Fellow of the Royal Geographical Society. He has received multiple international awards for research, teaching, and professional impact, including the ISOCARP Gerd Albers Award (2022, 2024), the Rising Scholar Award from the International Association for China Planning (2025), and the Early Career Teaching Award from the University of Hong Kong (2025). His work has also been recognized by the Forbes 30 Under 30 China list for Science and Healthcare. He serves as an Associate Editor for IET Smart Cities and the Journal of Urban Planning and Development, and regularly contributes to IEEE and interdisciplinary conferences and journals as a reviewer and committee member.

APPENDIX

A. LLM Prompts

System Prompt for Batch POI Name Matching ($C \rightarrow C'$ Mapping)

```
## Task
For each target store, find the best matching store from its corresponding nearby stores list.

Input:
- source_name: store name on platform A
- target_name: store name on platform B

Ignore punctuation, spacing, and case. Treat small bracket differences as the same.
Allow Chinese/English translations or pinyin equivalents if the brand and branch/location clearly match.

Return ONLY a JSON array with exactly one object, for example:
[{"matched_store":"XXX","reasoning":"short reason","confidence":95}]

Rules:
- If they are the same business: matched_store MUST equal target_name.
- If they are not the same business: matched_store MUST be "None".
- confidence is an integer 0-100 (higher = more confident they are the same shop).

## Example
**Input:**
```json
[{"id": "ex_001", "target_store": "McDonald's Restaurant", "nearby_stores": ["McDonald", "KFC", "Starbucks"]}]
```
Output:
```json
[{"id": "ex_001", "matched_store": "McDonald", "confidence": 95, "reasoning": "McDonald's Chinese name match"}]
```
```

User Prompt for Batch POI Name Matching ($C \rightarrow C'$ Mapping)

```
Consider the following data. For each target store, identify the best matching store from its list of nearby
stores, along with the reasoning and a confidence score: ```{input}```
```

B. VLM Prompts

System Prompt for Batch POI Logo / Storefront Images Matching

```
# Visual POI Consistency Matching Expert
You are an expert in visual POI matching. Your task is to determine if two images represent the same business
establishment by analyzing visual consistency.

Key criteria:
- Brand identity (logo/signage/colors)
- Storefront architecture
- Interior decoration
- Products/packaging
- Street/background context

Confidence (0-100):
100 = absolutely same shop
0 = absolutely different shops
If unclear → confidence=0, answer="no".

Return ONLY:
{"answer":"yes","confidence":85,"reasoning":"Short explanation"}

## Example
**Two images showing McDonald's storefront and Big Mac**
```json
{"answer": "yes", "confidence": "92"}
```

Two images showing different restaurant brands
```json
{"answer": "no", "confidence": "15"}
```
```

User Prompt for Batch POI Logo / Storefront Images Matching

```
Determine if the two input images depict the same shop or business, and return a confidence score.

<image1>  <image2>
```

C. Model Hyperparameters

All models were implemented in Python using `scikit-learn`, `XGBoost`, and `LightGBM`. Unless otherwise specified, we used each library's default settings. Input features were standardised using `StandardScaler`. The dataset was divided into an 8:1:1 split for training, validation, and testing. Specifically, we first separated 20% of the samples as the held-out test set (`test_size = 0.2`, `stratify = y`, `random_state = 42`), and then further partitioned the remaining data into a 90/10 split to obtain the validation subset used for threshold selection and model tuning. For each model, the decision threshold was selected on the test set by maximising the F1 score over thresholds in $[0, 1]$ with a step of 0.01.

(1) Logistic Regression: We used ℓ_2-regularised logistic regression with the following configuration:

- `C = 1.0` (inverse regularisation strength)
- `penalty = "l2"`
- `solver = "liblinear"`
- `max_iter = 1000`
- `random_state = 42`

(2) Random Forest: The random forest classifier was configured as:

- `n_estimators = 200` (number of trees)
- `max_depth = None` (nodes expanded until all leaves are pure or contain fewer than two samples)
- `random_state = 42`
- `n_jobs = -1` (use all available CPU cores)

(3) LightGBM: The LightGBM gradient boosting model was configured as:

- `n_estimators = 200`
- `learning_rate = 0.05`
- `max_depth = -1` (no explicit depth limit)
- `random_state = 42`
- `n_jobs = -1`
- `verbose = -1` (suppress training logs)

(4) XGBoost: The XGBoost classifier was configured as:

- `n_estimators = 200`
- `learning_rate = 0.05`
- `max_depth = 4`
- `subsample = 0.8` (row subsampling per tree)
- `colsample_bytree = 0.8` (feature subsampling per tree)
- `objective = "binary:logistic"`
- `eval_metric = "logloss"`
- `random_state = 42`
- `n_jobs = -1`

D. *Typical Low-confidence Cases*

ID	Crowdsourced review platform	Online delivery platform	Variables	ID	Crowdsourced review platform	Online delivery platform	Variables
1	Name: 洞庭小龙虾 	Name: 大碗里小龙虾（皮皮虾·大闸蟹·烤鱼） 	GD: 9.3m NSL = 0.8 NSB = 0.79 ASB = 0.93 VC = 0 p = 0.40 (unmatched√)	6	Name:新疆艾力烧烤店 	Name:新疆艾尼巴依羊肉串 	GD: 27.5m NSL = 0.8 NSB = 0.88 ASB = 0.93 VC = 0 p = 0.51 (matched×)
2	Name:钞记自选快餐 	Name:钞记河南烩面 	GD: 19.2m NSL = 0.8 NSB = 0.82 ASB = 0.93 VC = 0 p = 0.40 (unmatched√)	7	Name:杨婆婆重度老火锅 	Name:冒鲜锅·冒菜·毛肚牛肉（黄田店） 	GD: 12.7m NSL = 0.8 NSB = 0.86 ASB = 0.91 VC = 0 p = 0.55 (matched×)
3	Name:潘多拉粿条汤 	Name:粿然香潮汕粿条汤 	GD: 15.1m NSL = 0.8 NSB = 0.93 ASB = 0.92 VC = 0 p = 0.43 (unmatched×)	8	Name:洞庭轩 	Name:洞庭轩·本真湘味(1979店) 	GD: 18.7m NSL = 0.8 NSB = 0.80 ASB = 0.95 VC = 1 p = 0.55 (matched√)
4	Name:成都特色烧烤·羊肉串·小龙虾 	Name: 满分烧烤·海鲜·小龙虾（民治店） 	GD: 17.8m NSL = 0.8 NSB = 0.73 ASB = 0.98 VC = 0 p = 0.46 (unmatched√)	9	Name:牛家人大碗牛肉面 	Name:牛壹家大碗牛肉面.炸酱面 	GD: 18.5m NSL = 0.9 NSB = 0.93 ASB = 0.87 VC = 1 p = 0.57 (matched√)
5	Name:鲜卤肥肠·四川豆花 	Name:毕三爷•鲜卤肥肠•四川豆花(光明店) 	GD: 23.7m NSL = 0.85 NSB = 0.91 ASB = 0.91 VC = 1 p = 0.50 (matched√)	10	Name:橘鹅餐厅 	Name:橘鹅·盐焗鹅饭（台式卤肉饭·中式健康餐） 	GD: 9.3m NSL = 0.8 NSB = 0.80 ASB = 0.89 VC = 1 p = 0.60 (matched√)

Notes: These examples illustrate **typical low-confidence cases** produced by the preliminary fusion model when evaluated against the filtered candidate pool (C').
GD: spatial distance; **NSL**: LLM-based name similarity (GPT-4.1 Mini); **NSB**: BERT-based name similarity (cosine similarity); **ASB**: BERT-based address similarity (cosine similarity); **VC**: binary visual consistency inferred by the VLM (GPT-4.1 Mini); **p**: predicted probability that the CRP–ODP pair is a match. The predicted class (match/unmatched) is shown in parentheses. √ indicates agreement with manual annotation; × indicates disagreement.

TABLE A1: Typical Low-confidence Cases